Prediction of Length of Stay for Stroke Patients using Artificial Neural Networks

Abstract. Strokes are neurological events that affect a certain area of the brain. Since brain controls fundamental body activities, brain cell deterioration and dead can lead to serious disabilities and poor life quality. This makes strokes the leading cause of disabilities and mortality worldwide. Patients that suffer strokes are hospitalized in order to be submitted to surgery and receive recovery therapies. Thus, it's important to predict the length of stay for these patients, since it can be costly to them and their family, as well as to the medical institutions. The aim of this study is to make a prediction on the number of days of patients' hospital stays based on information available about the neurological event that happened, the patient's health status and surgery details. A neural network was put to test with three attribute subsets with different sizes. The best result was obtained with the subset with fewer features obtaining a RMSE and a MAE of 5.9451 and 4.6354, respectively.

Keywords: data mining; machine learning; artificial neural networks; stroke; length of stay

1 Introduction

It is known that the healthcare industry produces huge amounts of data every day that incorporates various sectors and areas of expertise. The information can go from hospital resources to the patients' health status and diagnosis of diseases [1]. Thus, the resultant data is represented in different types and formats, making it very heterogeneous [2]. The lack of structure and poor standard practices can often lead to a lack of quality in the produced healthcare data [3].

In a healthcare facility, data is collected and stored at a rapid pace, which promote the arising of Knowledge Discovery in Databases (KDD). KDD provide more in-depth knowledge obtaining hidden patterns that may exist in the collected data. Data mining (DM), the most important step of KDD, focus on the extraction of knowledge from large quantities of data, aiming at discovering important information to the industry [4]. Whereas a traditional data analysis performed by a human being is not possible, the application of Machine Learning algorithms can easily interpret the data and its details. Since machines can be taught how to properly look at data, their predictions often lead to low error values [5].

The application of Data Mining (DM) techniques on healthcare data brings a many advantage to the industry. Medical institutions can discover new and useful knowledge that otherwise would remain unknown [6]. By using these methods, the quality of the healthcare services can be improved, and new management rules can be implemented in order to increase the productivity. This also brings new ways of preventing fraud and

abuses that could highlight inappropriate patterns on insurance claims or medical prescriptions [2].

Strokes are a leading cause of long-term disabilities, poor quality of life and mortality worldwide. A patient that suffers a stroke can face permanent complications and psychological issues throughout their remaining life [7]. Depending on its intensity, a stroke frequently leads to the patient's death, making it one of the most recurrent epidemiology of this century. The aging of the population and unhealthy lifestyles increase its risks and unfavorable outcomes [8].

When the blood supply to the brain gets interrupted, the brain cells do not receive the needed blood flow for their normal function and start to die. This is called a brain attack, or a stroke [9]. The patients are hospitalized and submitted to surgeries depending on the type of the stroke. Therapeutic interventions aim to minimize the length of their hospitalization, since it can be costly both to the patient and their family, as well as to the hospital [10]. Thus, it becomes necessary to predict the length of stay (LOS) for stroke patients. It can depend of many factors, such as the stroke's intensity and the patient's health and recovery. Therefore, this is the goal of study: given some input data, with information related to this neurological event, patient and surgery details, create a model that predicts the LOS with the lowest possible error.

Artificial neural network (ANN) is the Machine Learning algorithm used on this study to predict the number of days of a stroke patient's hospital stay. These neural networks aim to mimic the function of the human brain, hence the use of biological designations such as neurons and synapses [11]. They differ from conventional algorithmic approaches in a way that they do not follow a set of instructions to solve a certain problem. They are often called "black boxes", since it is not possible to understand how they really work [12]. The truth is that neurons are very powerful units that can assume many roles in the storage of information, images recognition and classification problems [13].

Regarding the structure of this paper, it is divided in six major sections. After the current introduction, the second section presents related work. The methodology used is then described in next section. Section four describes the steps of the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology that was followed during the DM process. In section five the obtained results are presented and discussed. Finally, section six includes conclusions and future work.

2 Related Work

Researches associated to the prediction of LOS related to various specialty areas have become a relevant topic of study, since there are many advantages that the application of Machine Learning algorithms can bring to this area. The prior studies range from general medical divisions to specific medical diseases/treatments.

Hasanat et al. [14] aim to predict the LOS for patients in order to control the hospital costs and improve its efficiency. The researchers selected a subset of features using the information gain metric and tested it with various Machine Learning models. The

Bayesian network model obtained the best result for accuracy with a value of 81.28%, while the algorithm C4.5 resulted in a 77.1% of accuracy.

Lella et al. [15] created a novel prediction model for the LOS of patients admitted to hospitals. The Growing Neural Gas model obtained an accuracy value of 96.36% which was a best result than the ones produced by the likes of ZeroR, OneR, J48 and Self Organizing Map (SOM).

Combes et al. [16] presented an approach to estimate the LOS in an emergency department using models based on linear regression. The results were satisfactory, with an error of approximately 2 hours in 75% of cases.

Khajehali and Alizadeh [22] developed a study with the aim of explore the important factors affecting the LOS of patients with pneumonia in hospitals. This study concluded that Bayesian boosting method led to better results in identifying the factors affecting LOS (accuracy 95.17%).

Rezaianzadeh et.al. [23] carried out a study with the aim of determine the predictors of LOS in cardiologic care wards developed and carried out based on data-mining approaches. The median and mean LOS was 4 and 4.15 days, respectively. The factors associated with the increase in the LOS (more than 4 days) were: the ST segment elevation myocardial infarction (STEMI) diagnosis at the time of referral, being in the 50–70 years old group, history of smoking, high blood lipids, history of hypertension, hypertension at the time of admission, and high serum troponin levels.

Lee et. al. [24] attempted to identify potential predictors of intensive care unit (ICU) LOS (LOS) for single lung transplant patients. Several conclusions were obtained through this study, including: the median ICU LOS was 5 days, and this was highly correlated with the duration of mechanical ventilation; patients with pulmonary hypertension had the longest ICU LOS.

Machado et.al. [25] used real data to identify patterns in patients' profiles and surgical events, in order to predict if patients will need hospital care for a shorter or longer period of time after surgery for perforated peptic ulcer. The best accuracy obtained was 87.30% using JRip.

Silva et. al. [26] describe an implementation of a data mining project approach to predict the hospitalization period of cardiovascular accident patients. The best learning models were obtained by the IBk and Random Forest methods, which presented high accuracy values 91.47% and 88.16% respectively.

3 Methodology

For the development of this study, the methodology selected was the Cross-Industry Standard Process for Data Mining, commonly known as CRISP-DM. This process acknowledges six fundamental steps, namely: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment [17]. It is considered the most popular methodology for Data Mining projects, since it is very flexible and promotes transitions between the different phases.

The Machine Learning software Weka was used for the general analysis of the distribution and balance of the data. The data preparation tasks defined also resorted to this software. During this phase, many techniques were used, such as:

- One Hot Encoding, that consists in transforming all the categorical variables in binary attributes. Let's say that an attribute has 4 different classes. This technique turns the classes into different variables with values of 0 or 1. This way, the prediction models will be able to make better predictions [24];
- Attribute selection with WrapperSubsetEval (the algorithm chosen was Logistic Regression) in conjunction with the search method BestFirst (with the direction as forward);
- Attribute selection with CfsSubsetEval in conjunction with BestFirst (with the direction as backward).

On the other hand, the selection of attributes, as well as the implementation, training and testing of the ANN were made on the platform RStudio using the programming language R in order to induce the Data Mining Models (DMM).

4 Data Mining Process

4.1 Business Understanding

When stroke patients are admitted to the hospital after having suffered from a neurological event, information about their health status and lifestyle habits are recorded, as well as details about the stroke. After being submitted to surgery, the medical professionals register the needed information about the surgery that was performed and complications that the patient is feeling. Long stays at the hospital can be costly to the patients as well to the medical institutions. Therefore, healthcare institutions efficiency and productivity can be put at risk if the lengths of stay are unexpected.

Thus, the goal of this study is to help the medical professionals and management teams by aiming at predicting the LOS for stroke patients that are being admitted to hospitals.

4.2 Data Understanding

The data that was provided to this study was collected from a hospital located in Portugal. It is about patients that suffered from a stroke and had to be admitted to that hospital in order to be submitted to surgery. It includes various details about their health, such as the assessment of diabetes and smoking habits. It also reports on the characteristics of the neurological event that happened. Moreover, data about the surgery performed and any lingering complications is also contained on the dataset.

There were 36 attributes with 203 samples. From a first analysis, it was evident that some attributes were not relevant to the study at hand. Their reasons differ, but they offered no important information for the prediction process. The majority of the features presented percentages of missing values below 26%. The only exceptions were

two attributes that present the ankle-brachial index for both the right and left ankle. These two variables have 98% of values that are missing. The tasks that were done in order to properly prepare the data for the submission to the ML algorithms are described on the next section.

4.3 Data Preparation

The data preparation phase acts as a way of manipulating the data in order to obtain better results from the models. Thus, the data is submitted to various methods where the attributes' relevance for the study, the percentages of missing values, the derivation of new variables, the data transformation and the selection of attributes are all considered. This way, the data is properly prepared to be submitted to the Machine Learning algorithms. This step is especially important for the knowledge extraction process, because it can be the determining factor for a satisfactory prediction.

After a close inspection of the dataset, two attributes were identified as having no relevance for the end goal. The patient identification and the medical observations offer no meaningful information to the prediction process, since the former only serves as a way of identifying the patient and the latter is represented as a text box, where the medical professionals wrote down what they thought was needed. It has no specified structure.

The next task was the substitution of the class designations for incremental integers. It's easier for the model to process integers instead of actual words or entire sentences. This helps by reducing the computing time that a model takes to make a prediction on the given data.

Additionally, two new attributes were derived from existent ones in the dataset. The first variable is related to the number of days between the neurological event that the patient suffered and the hospital admission. The other one represents the length of the patients' hospital stay.

There are features that have high percentages of values that are missing. For this reason, they cannot be worked with, since it's impractical to use them for prediction proposes. Those variables are the index for the right and left brachial-ankle pressure.

On the other hand, any attribute that represented a date was also removed, since the new ones that were created offer more meaning to the Data Mining process.

The missing values that were present in the dataset were replaced with the mean and the mode of the classes for the numerical and categorical attributes, respectively. Moreover, the numerical variables were normalized as a way of having the results between 0 and 1. This normalization was performed using the equation (1). Also, the technique One Hot Encoding, introduced before, was applied to the dataset.

$$\text{Normalized value} = (X-Xmin)/(Xmax-Xmin) \quad (1)$$

Since the dataset contains many attributes, it's possible that there are variables that contribute more to the prediction process than others. For this reason, two attribute selection methods were applied. As a result of these techniques, a subset of attributes is selected, in which every attribute is individually evaluated by addressing its relevance to the prediction of the target variable chosen.

By the application of the attribute evaluator WrapperSubsetEval (the algorithm chosen was Logistic Regression) in conjunction with the search method BestFirst (with the direction as forward), 14 variables were selected.

Whereas, the evaluator CfsSubsetEval with BestFirst (with the direction as backward) selected just 7 variables. According to the Weka documentation [18], this evaluator assesses the worth of the features by considering their individual ability of prediction with the degree of redundancy between them. Both these evaluators and search methods are available in this software.

As follows, three use cases were defined, as presented on Table 1. The first corresponds to the original dataset with all the features, after the application of the data preparation techniques. The second use case is represented by the subset of attributes that were selected by the WrapperSubsetEval, while the third contains the variables produced by the evaluator CfsSubsetEval. These will be put to test in order to determine which set of attributes is more meaningful to the prediction process. The set that obtains the lowest error is then considered to be the most appropriate to predict correctly the target variable.

Table 1. Summary of the used datasets

Use Case	Number of attributes	Selection technique
1	33	-
2	14	WrapperSubsetEval
3	7	CfsSubsetEval

4.4 Modeling

The Machine Learning algorithm chosen to process the data and predict the desired results is ANN. This method allows scientists to make analysis on complex data [11]. Knowledge that was hidden can be extracted offering meaningful information to a business.

Many different configurations were tried out during the development of this study. However, the one that consistently obtained the best results was a neural network using the backpropagation algorithm. This is a very popular method where the output produced by the NN is evaluated in comparison to the correct output. When the results do not meet the expectations, the weights between the defined layers are modified. The process is then repeated until the error value is satisfactory [19].

The neural network created contained three hidden layers with different sizes. The first layer had 40 nodes, whereas the second included 20 and the third 10. The choice for the number nodes on each layer depends on the problem at hand. While different combinations were put to test, this composition registered the lowest errors and the least computing time. The learning rate was defined as 0.01, while the error threshold was 0.001.

These tunings of the NN allowed for a more customized modelling that had in mind the data provided and the problem that is to be solved.

4.5 Evaluation

The measures chosen to evaluate the performance of the prediction models created were Root Mean Squared Error (also known as RMSE) and Mean Absolute Error (abbreviated as MAE). Their definition is presented on Fig. 1 and Fig. 2, respectively.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Fig. 1. Definition of RMSE.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\, x_i - x \,|$$

Fig. 2. Definition of MAE

The first (**Erro! A origem da referência não foi encontrada.**), RMSE, is a rule that measures the average magnitude of the error. It is the square root of the average of squared differences between the prediction made and the desired result. It has the advantage of giving more importance to large errors [20]. On the other hand, MAE (**Erro! A origem da referência não foi encontrada.**) scores the average magnitude of the errors with no consideration for their actual direction. By removing the square vales of the error and considering its absolute value, the bias towards outlying points is removed [21]. The results obtained for these metrics are presented on the next section, as well as the analysis and discussion of the results.

5 Results and Discussion

Three different use cases were defined with the propose of identifying which attributes were more relevant to the prediction process. The results obtained for the performance metrics RMSE and MAE are shown in Table 2 for the defined use cases.

Table 2. RMSE and MAE values for the different use cases

Use Case	RMSE	MAE
1	6.2951	4.6350
2	7.6601	4.5478
3	5.9451	4.6354

In the first case, all the features after the application of the data preparation methods are submitted to the NN. This use case will serve as a comparison term between the whole set of features and different subsets with fewer attributes. This will result in an analysis that will determine if it is better to use smaller set of attributes or all of them.

Thus, the second study case contains only 14 variables. These were defined using an appropriate evaluator that selected the most relevant attributes for the prediction of the length of the hospital stay for stroke patients. Aiming at reducing even further the size of the subset of features, a third use case was created where only 7 attributes of the original dataset are included. These three distinct situations will enable a study on the different prediction capabilities of the models created using various sets of attributes as input data for the neural net.

Erro! A origem da referência não foi encontrada. shows that the third use case (the one with fewer attributes) obtained the best value for the first performance metric, RMSE. With a value of around 5.9451, it is the lowest RMSE recorded. Both the first use case as well as the second use case registered higher values for that measure. The first use case produced a result of approximately 6.2951, whilst the second use case resulted in the highest value (7.6601). This indicates that the smallest subset of attributes is a better input for the net, since it improves its predictions on the desired output. The set of all attributes also produce a better result than the subset of features that were selected by the evaluator WrapperSubsetEval.

Nonetheless, the results obtained for the MAE metric rank the models in a different way. The second use case produced a better result for this measure with a value of 4.5478. The other use cases did not obtain much worse results. In fact, the difference between them is not significant at all. The first use case had a MAE value of 4.6350 and the third a result of 4.6354. As can be seen, the different between the best and worst values is not bigger than 0.09.

As follows, since the divergence between the MAE values is not considered to be meaningful, the best set of attributes were selected by the CfsSubsetEval evaluator (third use case), which obtained the best RMSE result.

6 Conclusion and Future Work

Considered to be a dangerous epidemiology that is to be persistent on generations to come, strokes can lead to the death of patients. If not, the probability of suffering from physical complications and psychological issues is huge. Due to the interruption of the blood flow, important brain zones are put at risk of never recovering.

The application of Machine Learning algorithms can greatly improve the healthcare services performed, as well as increase the possibility of the patient's survival. Artificial neural networks aim to mimic the functioning of the human brain, making them an interesting model to prediction problems. In this research, given information about the patient's health, the stroke and the performed surgery, the goal was to produce a model capable of predicting the length of stay for stroke patients.

By testing three use cases with different sizes of feature sets, it was possible to define an optimal neural network configuration where the lowest error values were registered. It was concluded that the third use case, that is the one with fewer variables, obtained better results than the others attribute sets.

The future work includes getting more data detailing different aspects of the patient's health, stroke and surgery and test them with other neural network tunings.

References

1. Kautish, S., Abbas Ahmed, R.K.: A Comprehensive Review of Current and Future Applications of Data Mining in Medicine & Healthcare. International Journal of Engineering Trends and Technology, 38(2), 60–63 (2016).
2. Durairaj, M., Ranjani, V.: Data Mining Applications In Healthcare Sector: A Study. International Journal of Scientific & Technology Research, 2(10), 29-35 (2013).
3. Sukumar, S.R., Natarajan, R., Ferrell, R.K.: Quality of Big Data in health care. International Journal of Health Care Quality Assurance 28(6), 621-34 (2015).
4. Zhang, S., Zhang, C., Yang, Q.: Data preparation for data mining. Applied Artificial Intelligence, 17(5), 375-381 (2010).
5. Simeone, O.: A Very Brief Introduction to Machine Learning With Applications to Communication Systems. in IEEE Transactions on Cognitive Communications and Networking, vol. 4, no. 4, pp. 648-664, (2018).
6. Dennison, T., Qazi, F.: Data Mining in Health Care. In: Conference: Proceedings of The 2005 International Conference on Data Mining. pp: 89-9. CSREA Press, Las Vegas, Nevada, USA, (2005).
7. Clarke, D., Forster, A.: Improving post-stroke recovery: the role of the multidisciplinary health care team. Journal of multidisciplinary healthcare, 8, 433-442 (2015).
8. Dreyer, R., Murugiah, K., Nuti, S. V., Dharmarajan, K., Chen, S.I., Chen, R., Wayda, B., Ranasinghe, I.: Most Important Outcomes Research Papers on Stroke and Transient Ischemic Attack. Circulation: Cardiovascular Quality and Outcomes, 7(1), 191–204 (2014).
9. Gund, M.B., Jagtap, P.N., Ingale, V.B., Patil, R.Y.: Stroke: A Brain Attack. IOSR Journal Of Pharmacy, 3(8), 1-23 (2013).
10. Evrim, Gö., Turhan, K., Arzu, G., Vesile, Ö., Kursad, K.: The factors affecting length of stay in hospital among acute stroke patients. Journal of Neurological Sciences, 34(2), 143-152 (2017).
11. Maind, S.B., Wankar, P.: Research Paper on Basic of Artificial Neural Network, International Journal on Recent and Innovation Trends in Computing and Communication, 2, 96-100, (2014).
12. Benitez, J.M., Castro, J.L., Requena, I.: Are artificial neural networks black boxes? IEEE Transactions on Neural Networks, 8(5), 1156–1164 (1997).
13. Romaniuk, R.: Analysis of electrical patterns activity in artificial multi-stable neural networks. Proc. SPIE, vol.11176.
14. Al Taleb, A.R., Hoque, M., Hasanat, A., Khan, M.B.: Application of data mining techniques to predict length of stay of stroke patients. In: 2017 International Conference on Informatics, Health & Technology (ICIHT). pp. 1–5. IEEE (2017).
15. Lella, L., Di Giorgio, A., Dragoni, A.F.: Length of Stay Prediction and Analysis through a Growing Neural Gas Model. In: 2015 Artificial Intelligence and Assistive Medicine, pp. 11-21 (2015).
16. Combes, C., Kadri, F., Chaabane, S. Predicting hospital length of stay using regression models: Application to emergency department. In: 10éme Conference Francophone de Modélisation, Optimisation et Simulation- MOSIM'14 (2014).
17. Silva, E., Cardoso, L., Portela, F., Abelha, A., Santos, M.F., Machado, J.: Predicting Nosocomial Infection by Using Data Mining Technologies. In: Rocha A., Correia A., Costanzo S., Reis L. (eds) New Contributions in Information Systems and Technologies. Advances in Intelligent Systems and Computing, vol 354. Springer, Cham (2015).
18. More Data Mining with Weka, https://www.cs.waikato.ac.nz /ml/weka/mooc/moredataminingwithweka/, last accessed 2019/11/21

19. Skorpil, V., Stastny, J.: Neural Networks and Back Propagation Algorithm. In: ELECTRONICS 2016, Sozopol, Bulgaria, pp. 173-178 (2016).
20. Twomey, J.M., Smith, A.E.: Performance measures, consistency, and power for artificial neural network models. Mathematical and Computer Modelling, 21(1–2), 243-258 (1995).
21. Kneale, P., See, L., Smith, A.: Towards defining evaluation measures for neural network forecasting models. In: Proceedings of the Sixth International Conference on GeoComputation, University of Queensland, Australia (2001).
22. Khajehali, N., & Alizadeh, S.: Extract critical factors affecting the length of hospital stay of pneumonia patient by data mining (case study: an Iranian hospital). Artificial intelligence in medicine, 83, 2-13 (2017).
23. Lee, K. H., Martich, G. D., Boujoukos, A. J., Keenan, R. J., & Griffith, B. P.: Predicting ICU length of stay following single lung transplantation. Chest, 110(4), 1014-1017 (1996).
24. Rodríguez, P., Bautista, M. A., Gonzalez, J., & Escalera, S.: Beyond one-hot encoding: Lower dimensional target embedding. Image and Vision Computing, 75, 21-31 (2018).
25. Machado, J., Cardoso, A. C., Gomes, I., Silva, I., Lopes, V., Peixoto, H., & Abelha, A.: Predicting the Length of Hospital Stay After Surgery for Perforated Peptic Ulcer. In International Conference on Information Technology & Systems, pp. 569-579. Springer, Cham (2019).
26. Silva, C., Oliveira, D., Peixoto, H., Machado, J., & Abelha, A.: Data Mining for Prediction of Length of Stay of Cardiovascular Accident Inpatients. In International Conference on Digital Transformation and Global Society, pp. 516-527. Springer, Cham (2018).